Using confidence intervals for assessing reliability of real tests

Oosterwijk, P.R.; van der Ark, L.A.; Sijtsma, K.

Published in:
Assessment

DOI:
10.1177/1073191117737375

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Article

Using Confidence Intervals for Assessing Reliability of Real Tests

Assessment
1–10

DOI: 10.1177/1073191117737375
journals.sagepub.com/home/asm
SAGE

Pieter R. Oosterwijk[1], L. Andries van der Ark[2], and Klaas Sijtsma[3]

Abstract

Test authors report sample reliability values but rarely consider the sampling error and related confidence intervals. This study investigated the truth of this conjecture for 116 tests with 1,024 reliability estimates (105 pertaining to test batteries and 919 to tests measuring a single attribute) obtained from an online database. Based on 90% confidence intervals, approximately 20% of the initial quality assessments had to be downgraded. For 95% confidence intervals, the percentage was approximately 23%. The results demonstrated that reported reliability values cannot be trusted without considering their estimation precision.

Keywords

confidence intervals for reliability, precision of reported reliability, quality assessment of reliability, test-score reliability

General Introduction

Scores on psychological tests and questionnaires are used for making high-stakes decisions about hiring applicants for a job or rejecting them; assigning or withholding a patient a particular treatment, a therapy, or a training; accepting students at a school or rejecting them; enrolling students in a course or rejecting them; or passing or failing an exam. In these applications, the stakes are high for the individuals and the organizations involved, and tests must satisfy a couple of quality criteria to guarantee correct decisions. For example, the test score must be highly reliable and valid, and norms must be available to interpret individual test performance. In lower-stakes applications, tests also must satisfy quality requirements but usually lower than for high-stakes decisions (Evers, Lucassen, Meijer, & Sijtsma, 2010). For example, an inventory may be used to assess personal interests to help clarifying the kind of follow-up education a high school student might pursue. Often, the inventory is only one of the many data sources used next to, for example, school and parental advice. Another example is the use of the test score as the dependent variables in experiments (e.g., degree of anxiety) or an independent variable in linear explanatory models (e.g., as predictors of therapy success).

The assessment of a test involves many different quality aspects (Clark & Watson, 1995). This study focuses on test-score reliability. In particular, we study the problem that in test construction research, test constructors tend not to estimate confidence intervals (CIs) for test-score reliability and thus do not take the uncertainty of the estimates into account when assessing the quality of their test (Fan & Thompson, 2001). For example, a sample reliability equal to .84 is incorrectly treated as if it were a parameter not liable to sampling error and it is concluded, for example, that a test has a reliability of .84, ignoring that a 95% CI equal to, say, (.74; .91), would suggest true reliability may be considerably higher or lower than .84. Kelley and Cheng (2012) argued that CIs may be more important than reliability point estimates, and Wilkinson and the Task Force on Statistical Inference (1999) provided general guidelines for the use of statistics such as CIs in psychological research. In addition, test assessment agencies tend to base their assessments of reliability on the estimate thus ignoring sampling error (e.g., Evers, Lucassen, et al., 2010). This means that if they consider reliability denoted by ρ in excess of a criterion value of, say, c, to be "good," they make the decision provided sample reliability $r > c$ without statistically testing whether $\rho > c$ given sample value r.

Maxwell, Kelley, and Rausch (2008) emphasized the importance of sample size considerations to obtain CIs allowing simultaneously to assess the direction, the magnitude (the authors refer to estimation precision), and the accuracy of an effect. For reliability, this translates to

[1]Court of Audit, The Hague, Netherlands
[2]University of Amsterdam, Amsterdam, Netherlands
[3]Tilburg University, Tilburg, Netherlands

Corresponding Author:
L. Andries van der Ark, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands.
Email: L.A.vanderArk@uva.nl

assessing whether, based on the available sample, one has enough evidence that $\rho > c$ (c must not be in the CI), which is a power issue, whether one can pinpoint ρ to a sufficiently narrow range of plausible values, and whether one can be confident that an estimate of ρ is unbiased. The latter topic is problematic in reliability estimation, because all available methods are known to be lower bounds, and hence negatively biased, but it is also known which ones are more accurate, however leaving the magnitude of the bias unknown. Oosterwijk, Van der Ark, and Sijtsma (2017) discuss estimation procedures that tend to be positively biased. In this study, direction and precision are most relevant, and using sample r rather than the CI for ρ to make decisions invites reliability assessments that are too optimistic, providing test practitioners, their clients and patients with measurement instruments that promise better psychometric quality than is realistic, a situation one should want to avoid.

Relevance of the Study

We estimated the magnitude of the problem of ignoring CIs and treating reliability estimates as if they were parameter values in practical test construction and test quality assessment. We investigated this in a large database in which test assessments are collected (Egberink, Janssen, & Vermeulen, 2009-2016). The database is operated by the Dutch Committee on Tests and Testing (acronym COTAN) that works under the auspices of the Dutch Association of Psychologists (acronym NIP). Dutch and Dutch-language Belgian test constructors and test practitioners appreciate COTAN to assess their tests and the results published in the database. We investigated to what degree not taking CIs into account and relying solely on reliability point estimates affected the assessments of tests' reliability. We determined the percentage of tests in the database for which we had to change the quality assessment when CIs were considered instead of point estimates.

COTAN is an active test assessment agency of good reputation that has assessed the quality of tests and questionnaires since 1959; also see Evers, Sijtsma, Meijer, and Lucassen (2010) and Sijtsma (2012). Dutch governmental and insurance companies require COTAN's approval of tests as a necessary condition for accepting requests for particular benefits and payments, respectively. For the majority of the tests in the database, statistical information needed to estimate CIs was unavailable and despite great effort we were able to retrieve only little additional information from university libraries. Incompleteness of the available subset of tests concerns a typical problem found in meta-analysis, possibly introducing bias in the results. Despite this drawback, we expect we can have more confidence in tests for which complete information was available than in tests for which information was lacking. In addition, the available tests represent various psychological attributes well, thus sufficiently covering the testing field. The widespread use of tests in the Netherlands guarantees some degree of generality of the results, thus mitigating the call for a sample of tests from a larger geographic region. This study is unique and the available test subset is comprehensive even though it is incomplete.

Based on a sample estimate of the test-score reliability the test constructor reports, and using a generally accepted classification system that we discuss later, the COTAN database classifies the tests' reliability as insufficient, sufficient, or good. Dutch and Belgian test constructors accept and use the COTAN classification system for test assessment including the reliability classification, as a guide for test construction, which amplifies its importance even though the classification is arbitrary to some extent and other guidelines are available in the literature. For different reliability classifications, see Nunnally (1978, p. 246), Cascio (1991, p. 141), Clark and Watson (1995), Murphy and Davidshofer (1998, pp. 142-143), DeVellis (2003, pp. 94-95), Smith and Smith (2005, pp. 121-122), Gregory (2007, p. 113), and McIntire and Miller (2007, p. 202). Evers et al. (2013, pp. 43-52), on behalf of the European Federation of Psychologists' Associations (EFPA), provided four categories for reliability, the highest of which was labelled "Excellent" for high-stakes testing ($r \geq .9$) and the next "Good" ($.8 \leq .9$). Christensen (1997, pp. 217-219) discussed recommendations for reliability for dependent variables in experiments.

We chose to investigate the reliability rather the standard error of measurement, although one might argue that the latter quantity should be preferred for assessing the quality of decisions about individuals on the basis of test scores (e.g., Mellenbergh, 1996). Because the standard error of measurement is based directly on reliability, the choice for either one is arbitrary. Moreover, researchers routinely report reliability (e.g., AERA, APA, & NCME, 2014; Wilkinson and the Task Force on Statistical Inference, 1999), and test agencies assess reliability prior to the standard error of measurement, emphasizing reliability's pivotal position in measurement assessment.

Based on our experience, we had no knowledge of articles reporting CIs for reliability and a quick and modest literature scan did not alter this conclusion. We found this absence remarkable, because in particular for coefficient alpha (Cronbach, 1951) methods for estimating standard errors and CIs have long been available (e.g., Feldt, 1965; Feldt, Woodruff, & Salih, 1987; Hakstian & Whalen, 1976; Kristof, 1963). In addition, several authorities have urged researchers to report CIs (e.g., AERA, APA, & NCME, 2014), but apparently so far this has had little success.

Organization of the Article

This article is organized as follows. First, the vast majority of the tests we studied used coefficient alpha (e.g., Cronbach,

1951) and a non-ignorable minority used the split-half method (e.g., Lord & Novick, 1968, pp. 135-136). Other methods were rarely used. Because split-half method and coefficient alpha are based on classical test theory (Lord & Novick, 1968), we discussed reliability as defined by classical test theory, and split-half reliability and coefficient alpha. For both methods, we showed how CIs can be computed. Sijtsma and Van der Ark (2015) discuss other approaches based on factor analysis and generalizability theory. Second, we discuss collecting reliability data for this study from the online database of COTAN that is available to paid subscribers, and we discuss both the assessment of reliability standards without (i.e., COTAN) and with (i.e., our approach) using CIs. Third, we present the results of the reliability data collection from the COTAN online database and we discuss the reliability assessment results using CIs for reliability and compare the results with the assessments COTAN published. Finally, we outline the results of this study and their meaning for future reliability assessment.

Reliability and Estimation Methods

Classical Test Theory and Definition of Reliability

Assume that a psychological test consists of J items indexed by j ($j = 1,\ldots,J$). Let variable X_j denote the score on item j. The test score is the sum of item scores X_j, defined as $X = \sum_{j=1}^{J} X_j$, with population variance, σ_X^2. Classical test theory assumes that X is the sum of an unobservable true score T and an unobservable random measurement error E, with variances σ_T^2 and σ_E^2. Because random measurement error E is assumed to be uncorrelated with true score T, the variance of the test score can be decomposed as $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Two tests with test scores X and X' are parallel if (1) for each person i his true scores must be equal, $T_i = T_i'$, implying that in the group $\sigma_T^2 = \sigma_{T'}^2$, and (2) the variance of the test scores in the group must be equal, $\sigma_X^2 = \sigma_{X'}^2$. The reliability of the test score is defined as the product–moment correlation of X and X', denoted $\rho_{XX'}$, and it is well-known that (Lord & Novick, 1968, p. 61)

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T'}^2}{\sigma_{X'}^2}. \quad (1)$$

Reliability ranges from 0 (if $\sigma_T^2 = 0$) to 1 (if $\sigma_T^2 = \sigma_X^2$, meaning $\sigma_E^2 = 0$). In practice, reliability is almost never 1, but several tests from the COTAN database had high reliability even up to 1.00 (results available on request from the authors). Reliability estimates lower than, say, $.6$, were rarely reported. For a factor-analysis approach to reliability, Markon and Chmielewski (2013) discuss how model misspecification can cause reliability estimates to be outside the $[0;1]$ interval.

Reliability in Equation (1) cannot be computed in practice, because parallel test scores X and X' are rarely available, and both true-score variances σ_T^2 and $\sigma_{T'}^2$ are unobservable. In practical test research, usually one has data available from one test and one test administration, and several methods have been proposed to estimate reliability in this situation (e.g., Bentler & Woodward, 1980; Cronbach, 1951; Guttman, 1945; Lord & Novick, 1968; Ten Berge & Zegers, 1978; Zinbarg, Revelle, Yovel, & Li, 2005).

The Two Reliability Methods Used in the COTAN Database

We investigated the split-half method and coefficient alpha. Cronbach (1951) argued that the latter method must replace the former, an advice that test constructors took to heart, making the easy to use coefficient alpha by far the most popular reliability method (Heiser et al., 2016). Despite heavy criticism (e.g., Clark & Watson, 1995; Cortina, 1993; Cronbach & Shavelson, 2004; Schmitt,1996; Sijtsma, 2009) and the existence of many alternatives providing better approximations to reliability (Sijtsma & Van der Ark, 2015), coefficient alpha continues to be the reliability method most frequently used (Heiser et al., 2016). We limited our attention to the split-half method and coefficient alpha, not because we prefer these methods but because they are predominantly used in the COTAN database.

Split-Half Method. The researcher splits his test in two halves, correlates the test scores obtained on the halves, and uses a correction formula to obtain an estimate of the reliability for the whole test. Formally, two situations may be distinguished. First, when the test halves are parallel, the product–moment correlation between the half-test scores Y_1 and Y_2, denoted $\rho_{Y_1Y_2}$, by definition equals the reliability of the test score on a half test, $\rho_{YY'}$; that is, $\rho_{Y_1Y_2} = \rho_{YY'}$. Then, the reliability of the test score on the complete test, $\rho_{XX'}$, can be computed by means of the Spearman–Brown prophesy formula (Lord & Novick, 1968, p. 84) adapted to doubling test length,

$$\rho_{XX'} = \frac{2\rho_{YY'}}{1+\rho_{YY'}}. \quad (2)$$

Second, when test halves are not parallel, Equation (2) produces an invalid result; that is, $\rho_{Y_1Y_2} \neq \rho_{YY'}$ and inserting $\rho_{Y_1Y_2}$ does not produce reliability $\rho_{XX'}$ but a value that one may denote as SH, for which $SH \neq \rho_{XX'}$.

Methods to compute a CI for *SH* are available (Charter, 2000; Fan & Thompson 2001). Let $r_{Y_1Y_2}$ denote the sample correlation between the two test scores on the test halves. A CI for *SH* can be constructed that takes the asymmetrical

sampling distribution of the product–moment correlation into account. First, the estimate of the correlation between two test halves ($r_{Y_1Y_2}$) is obtained. Second, $r_{Y_1Y_2}$ is transformed using the Fisher Z transformation,

$$Z = .5\ln\left(\frac{1+r_{Y_1Y_2}}{1-r_{Y_1Y_2}}\right). \quad (3)$$

Z is approximately normally distributed with a mean equal to $.5\ln\left(\frac{1+\rho_{Y_1Y_2}}{1-\rho_{Y_1Y_2}}\right)$ and a standard error approximately equal to $\frac{1}{\sqrt{N-3}}$ (e.g., Hays, 1994, p. 649). Third, let α denote the nominal Type I error rate. Let ζ be the parameter corresponding to Z; and let $Z_{\alpha/2}$ be the lower bound and let $Z_{1-\alpha/2}$ be the upper bound of a $(1-\alpha)$ 100% CI for ζ. For a 95% CI, the lower bound equals $Z_{\alpha/2} = Z - \frac{1.96}{\sqrt{N-3}}$ and the upper bound equals $Z_{1-\alpha/2} = Z + \frac{1.96}{\sqrt{N-3}}$, so that the 95% CI equals $(Z_{\alpha/2}; Z_{1-\alpha/2})$ or, equivalently,

$$\left(Z - \frac{1.96}{\sqrt{N-3}}; Z + \frac{1.96}{\sqrt{N-3}}\right). \quad (4)$$

Fourth, the bounds of the CI can be transformed into bounds on the $r_{Y_1Y_2}$ scale using the inverse of Equation (3),

$$r_{Y_1Y_2} = \frac{e^{2Z}-1}{e^{2Z}+1}. \quad (5)$$

Finally, after having obtained the bounds of a CI for $r_{Y_1Y_2}$, Equation (2) is used to transform the bounds into bounds on the SH scale. The resulting CI is asymmetrical.

Coefficient Alpha. Let the covariance between items j and k be denoted σ_{jk}; then, coefficient alpha is defined as

$$\text{alpha} = \frac{J}{J-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sigma_X^2}. \quad (6)$$

Given classical test theory assumptions, alpha is a lower bound to the reliability; $\text{alpha} \leq \rho_{XX'}$ (Novick & Lewis, 1967). Other authors (e.g., Bentler, 2009) have noted that alpha can also overestimate reliability when a factor analysis approach to reliability is pursued, but test constructors of the tests we assessed did not follow this approach. Standard errors for the sample estimate $\widehat{\text{alpha}}$ and CIs for alpha have been derived (e.g., Feldt, 1965; Feldt et al., 1987; Kuijpers, Van der Ark, & Croon, 2013; Maydeu-Olivares, Coffman, & Hartmann, 2007; Van Zyl, Neudecker, & Nel, 2000). The standard errors these authors proposed to estimate CIs for alpha produce symmetrical intervals whereas alpha is bounded from above by the value 1.

In this study, we used Feldt's method (Feldt et al., 1987). Feldt's method is convenient because it uses only information available for several tests in the COTAN database: $\widehat{\text{alpha}}$, test length J, and sample size N. A drawback is that failure of the method's assumptions may bias standard errors and CIs, especially as alpha values are higher (Kuijpers et al., 2013). Higher alpha values are the more interesting values in our study, and using alternative but mathematically more involved methods for determining standard errors that address the bias problem might have solved this problem, were it not that such methods require the availability of statistical information that test manuals usually did not report, thus rendering the use of these methods impossible. For examples of more involved methods, see Maydeu-Olivares et al. (2007), Kelley and Cheng (2012), and Kuijpers, et al. (2013).

To compute the 95% CI for alpha, let the nominal Type I error rate be 0.05, and let F_a and F_b be critical values of an F distribution with $N-1$ and $(N-1)(J-1)$ degrees of freedom, such that $P(F < F_a) = .025$ and $P(F < F_b) = .975$. For example, using Hays (1994, pp. 1016-1022) for $N = 100$ and $J = 10$, one finds that $F_a \approx 0.7315$ and $F_b \approx 1.3198$. Feldt et al. (1987) showed that the 95% CI for alpha is estimated by

$$\left(1-\left[1-\widehat{\text{alpha}}\right]\times F_b; 1-\left[1-\widehat{\text{alpha}}\right]\times F_a\right). \quad (7)$$

Method

We used the COTAN database to answer two questions: (1) What is the precision of reported reliability estimates expressed by 90% and 95% CIs; (2) Does considering precision change the qualification tests initially received with respect to reliability?

Test Population and Test Sample

COTAN assesses the most recent versions of tests that are used in the Dutch and Belgian practice for testing individuals to obtain a diagnosis, give an advice, or make a decision, and in addition COTAN assesses tests used in scientific research. COTAN's database distinguishes three main test types: (1) person–situation tests measuring proficiency in a particular setting, such as employment or education. Examples are questionnaires assessing people's vocational interests, tests for school achievement, but also inventories assessing patients' behavior in mental institutions; (2) person tests measuring personality, addressing stable personality traits such as the big five, and also intelligence; and (3) situation tests assessing situational performance, which concern, for example, expert ratings of labor situations' task characteristics and students' judgments of the affective meaning of concepts. These three main test types are subdivided into 38 finer grained test types, which however are

Table 1. Tests Included in the Analysis and Tests Excluded From the Analyses, Arranged by Assessment Category and Test Type.

	Included	Excluded	Total
Assessment			
Insufficient	18 (17.3%)	120 (83.7%)	138 (100%)
Sufficient	57 (26.3%)	160 (73.7%)	217 (100%)
Good	41 (24.8%)	124 (75.2%)	165 (100%)
Total	116 (22.3%)	404 (77.7%)	520 (100%)
Test type			
Person-Situation	55 (17.8%)	254 (88.2%)	309 (100%)
Person	34 (22.2%)	119 (77.8%)	153 (100%)
Situation	12 (66.7%)	6 (33.3%)	18 (100%)
Two types	15 (37.5%)	25 (62.5%)	40 (100%)
Total	116 (22.3%)	404 (77.7%)	520 (100%)

Note. Included = tests for which at least one CI (confidence interval) could be estimated; Excluded = tests for which number of items, sample size, or reliability were not reported, including 51 tests for which no reliability research whatsoever was reported.

not useful for our study. Using an assessment system created by COTAN (Evers, Lucassen, et al., 2010), raters commissioned by COTAN assessed the reliability of 520 tests to have insufficient (138 tests), sufficient (217), or good (165) reliability. The tests were 309 person–situation tests, 153 person tests, and 18 situation tests. Forty tests may be placed in more than one category.

Collecting Tests and Composition of Test Subset

We distinguish test batteries and single tests. Test batteries consist of several subtests, test scores being provided for subtests and for the whole battery based on the subtest scores. Single tests measure one attribute and are either subtests from test batteries or tests measuring one attribute that are not part of a test battery.

Test publishers provide COTAN with a copy of the test and all corresponding materials including the manual, but COTAN is not allowed to grant researchers, like the present authors, access to these materials. Hence, we retrieved more detailed information from the COTAN online database (Egberink et al., 2009-2016) and test manuals available from libraries of the University of Amsterdam and Tilburg University.

To compute a CI, one needs number of items (*J*), sample size (*N*), and reliability estimate (*r*). For 116 (22.3%) out of 520 tests COTAN assessed the results were complete, hence we discarded 404 tests from the analysis for which *J, N*, or *r* were missing. In Table 1, entry "41" (Table 1; 3rd row, 1st column) should be read as "For 41 tests COTAN assessed to have "Good" reliability, we could retrieve all the relevant results." These 41 tests entail both test batteries that are counted once, also when they were assessed for different

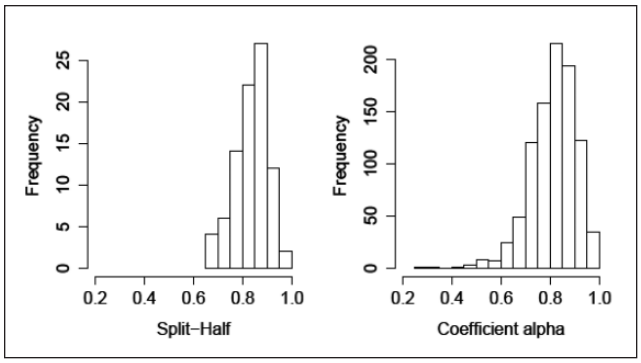

Figure 1. Split-half reliability (87 estimates) and coefficient alpha (937 estimates) distributions.

groups, and single tests. Comparing categories for included and excluded tests, Table 1 shows that percentages vary little across assessment categories and test types (except for Situation tests, but here the frequencies were small), suggesting absence of bias due to lack of representation.

The 116 tests produced 1,024 reliability estimates, 105 of which pertain to total scores on a test battery and 919 to single tests. Most reliability estimates (74.71%) were based on at most 20 items, and four tests contained more than 200 items. More than half of the reliabilities were estimated from samples smaller than 1,000 observations, and 53 reliabilities were estimated from samples ranging from 6,294 to 12,522 observations. Most (94.73%) reliability estimates varied between 0.60 and 0.95 (Figure 1). The split-half method was reported 87 times and coefficient alpha 937 times.

Frequencies *M* in Table 2 count the number of reliability values retrieved for test batteries and single tests, arranged by quality assessment and test type. Tables 1 and 2 are related as follows. The 41 tests enumerated in Table 1 (3rd row, 1st column) produce 55 reliability values (Table 2; 3rd row, 1st column) for total scores on test batteries, also separately counting available subgroup results; and 468 reliability values (3rd row, 5th column) based on single tests. For each count *M* mean number of items, sample size, and reliability are provided.

Reliability Assessment Rules

COTAN Rules. COTAN distinguishes three mutually exclusive and exhaustive reliability intervals labelled "Insufficient" (I), "Sufficient" (S), and "Good" (G) to assess reliability. Let r denote a reported, estimated reliability, let c_{IS} denote the reliability value that separates "Insufficient" from "Sufficient," and c_{SG} the reliability value that separates "Sufficient" from "Good." Hence, the three regions are defined by $(0; c_{IS}]$ for "Insufficient"; $(c_{IS}; c_{SG}]$ for "Sufficient"; and $(c_{SG}; 1)$ for "Good." COTAN assessments

Table 2. Descriptive Statistics for Reliability Estimates, Separately for Test Batteries and Single Tests, Arranged by Assessment Category and Test Type.

	Test battery				Single test			
	M	$\bar{J}$	$\bar{N}$	$\bar{r}_{XX'}$	*M*	$\bar{J}$	$\bar{N}$	$\bar{r}_{XX'}$
Assessment								
Insufficient	16	25.56	530.06	.71	142	11.94	844.28	.68
Sufficient	34	34.65	550.50	.83	309	12.63	1611.60	.79
Good	55	63.16	863.29	.90	468	19.10	1338.39	.88
Test type								
Person–situation	52	57.63	722.75	.85	483	12.68	1910.52	.80
Person	41	41.56	627.10	.84	254	18.47	844.16	.83
Situation	9	31.44	1103.11	.86	118	16.51	628.47	.82
Two types	3	25.67	485.67	.89	64	23.91	513.83	.88
Total	105	48.20	711.23	.85	919	15.71	1353.91	.82

Note. *M* = number of reliability estimates; $\bar{J}$ = mean number of items used for computing coefficient alpha; $\bar{N}$ = mean sample size used for computing reliability estimate; $\bar{r}_{XX'}$ = mean reported reliability estimate.

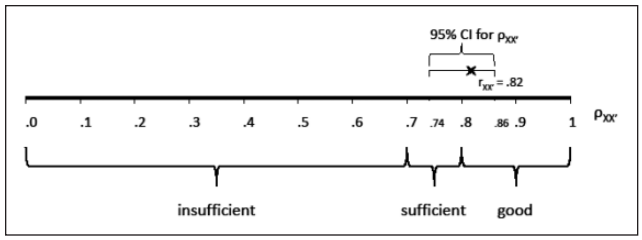

Figure 2. Example of qualification of reliability and CI for reliability.

are formalized as follows: if $r \in (0; c_{IS}]$, then assign "Insufficient"; if $r \in (c_{IS}; c_{SG}]$, then assign "Sufficient"; and if $r \in (c_{SG}; 1)$, then assign "Good."

COTAN distinguishes three different uses of tests, which are, in decreasing order of importance reflected by smaller c_{IS} and c_{SG} values: (1) making important decisions about individuals, such as admittance to a school or selection for a job ($c_{IS} = .8$ and $c_{SG} = .9$); (2) obtaining an impression about an individual's personality to help that individual think about the kinds of jobs he might consider pursuing after he/she has completed school ($c_{IS} = .7$ and $c_{SG} = .8$); and (3) using the test score for group-level measurement, for example, in a research project that studies differences between the arithmetic skills in different age groups ($c_{IS} = .6$ and $c_{SG} = .7$).

Confidence Intervals. For individual advice, Figure 2 presents a numerical example for $c_{IS} = .7$ and $c_{SG} = .8$, and a test for which $r = .82$. Following COTAN decision rules, $.82 \in [.8; 1]$; hence, assign "Good." Assume that CI equals (.74; .86); then, because $c_{SG} \in (.74; .86)$, r is not significantly larger than c_{SG} so that "Good" is ruled out but "Insufficient" and "Sufficient" are open. Next, $c_{IS} \notin (.74; .86)$; hence, r is significantly larger than c_{IS} and "Sufficient" is assigned for this reliability value ("Insufficient" is ruled out). We considered 90% and 95% CIs, implying nominal one-sided Type I errors of 0.05 and 0.025, respectively, for the test that a reliability value is significantly greater than a lower threshold value.

Let L denote the lower bound of the CI and U the upper bound. The formalized decision rule taking CIs into account is as follows: (1) if $r < c_{IS}$, then assign "Insufficient"; (2) if $c_{IS} \le r < c_{SG}$, then determine if $c_{IS} \in (L; U)$; if so, then assign "Insufficient," else assign "Sufficient"; (3) if $r \ge c_{SG}$, then determine if $c_{IS} \in (L; U)$; if so, then assign "Insufficient"; else determine if $c_{SG} \in (L; U)$; if so, then assign "Sufficient," else assign "Good."

The decision rule that takes CIs into account cannot upgrade a reliability value to a higher category, because it tests whether a sample reliability value is significantly larger than a cut-off score; if yes, the original COTAN assignment is maintained, else it is downgraded. We chose our somewhat conservative procedure to protect the test practitioner and his clients and patients from tests that provide less quality than the assessment promises.

Results

As test batteries and single tests are longer and samples are larger, reliability and its assessment increase (Table 2) and mean $CI_{95\%}$ and $CI_{90\%}$ width decreases (Table 3), implying greater statistical certainty. Compared with test batteries, single tests contain fewer items, are based on larger samples, and have lower reliability, but mean CI width is approximately equal. For test type, Table 2 shows for test batteries that person–situation tests on average are the longest and are based on the largest samples. Different test

Table 3. Proportion of Reliability Estimates (*Pr*) for Which the Lower Bound of the 90% CIs and 95% CIs Exceeds the c_{IS} (Sufficient Category) and c_{SG} (Good Category) Lower Bounds, Arranged by Assessment Categories and Test Types.

	Test battery					Single tests				
		95% CI		90% CI			95% CI		90% CI	
	M	*W*	*Pr*	*W*	*Pr*	*M*	*W*	*Pr*	*W*	*Pr*
Qualification										
Insufficient	16	.09	NA	.08	NA	142	.09	NA	.07	NA
Sufficient	34	.05	.71	.04	.79	309	.06	.67	.04	.71
Good	55	.03	.82	.03	.84	468	.04	.76	.02	.79
Test type										
Person–situation	52	.04	.83	.03	.85	483	.06	.66	.04	.69
Person	41	.06	.56	.05	.63	254	.05	.74	.04	.78
Situation	9	.05	.89	.04	.89	118	.05	.69	.04	.74
Multiple types	3	.04	1.0	.03	1.0	64	.04	.80	.03	.84
Total	105	.05	.73	.04	.77	919	.06	.70	.04	.73

Note. CI = confidence interval; *M* = number of reliability estimates; *W* = mean CI width; *Pr* = proportion of reliability estimates that need not be downgraded; NA = not available ("Insufficient" assessment category does not have lower boundary).

types have almost the same mean reliability. Person–situation tests have the smallest mean CI width.

Table 3 shows the proportions of reliability estimates that need not be downgraded when taking confidence intervals into account (*Pr*); that is, reliability estimates for which the lower bound of the 90% CIs and 95% CIs exceeds the c_{IS} lower bound of the "Sufficient" category and the c_{SG} lower bound of the "Good" category. Proportions for $CI_{90\%}$ by definition are larger than for $CI_{95\%}$. The "Insufficient" category does not have a lower boundary; hence, *Pr* is not available (NA); 67% to 79% of the tests with the qualification "Sufficient" exceeded the c_{IS} threshold, and 76% to 84% of the test with the qualification "Good" exceeded the c_{SG} threshold. For person–situation test batteries and situation test batteries (lower-left panel), CI lower bounds exceeded *c* thresholds more often than for person test batteries.

Turnover Table 4 shows the COTAN assessments in the columns and the assessment based on CIs in the rows, with blanks in the lower triangles because using CIs can only produce the same or a lower assessment. Diagonal entries show frequencies of reliability estimates that were not reclassified. For test batteries, using 90% CIs, the entries add up to 16 + 27 + 46 = 89 (84.8% of 105 test batteries). Using 95% CIs, 81.0% of the reliability estimates were not reclassified. Of the 34 reliability estimates that were initially classified as "Sufficient," 20.6% (90% CIs) and 29.4% (95% CIs) were reclassified as "Insufficient." Of the 55 reliability estimates initially classified as "Good" 16.4% (90% CIs) and 18.2% (95% CIs) were reclassified to "Sufficient," and in both cases none were reclassified as "Insufficient." For single tests, out of 919 tests, 79.3% (90% CIs) and 76.5% (95% CIs) were not reclassified. Of the 309 reliability estimates originally classified as "Sufficient," 29.1% (90% CIs) and 33.3% (95% CIs) were reclassified as "Insufficient." Of the 468 reliability estimates originally classified as "Good," 20.7% (90% CIs) and 23.5% (95% CIs) were reclassified as "Sufficient," and 0.6% (90% CIs and 95% CIs) were reclassified as "Insufficient."

Table 4. Turnover Results for Assessment Categories for Test Batteries (Upper Panel) and Single Tests (Lower Panel), Without CIs and Using 90% and 95% CIs.

		Without CIs			
		I	S	G	Total
Using 90% CIs	I	16	7	0	23
	S		27	9	36
	G			46	46
Using 95% CIs	I	16	10	0	27
	S		24	10	34
	G			45	45
	Total	16	34	55	105
Using 90% CIs	I	142	90	3	235
	S		219	97	316
	G			368	368
Using 95% CIs	I	142	103	3	248
	S		206	110	316
	G			355	355
	Total	142	309	468	919

Note. CI = confidence interval; Without CIs = Qualification of reliability estimates using COTAN standards; Using 90% CIs = qualification of the reliability estimates using 90% CIs; Using 95% CIs = qualification of the reliability estimates using 95% CIs; I = insufficient; S = sufficient; G = good.

Discussion

Using CIs for test batteries, almost 20% of the reliability estimates had to be downgraded to the next lower category, and for single tests the percentage exceeded 20%, but downgrading from "Good" to "Insufficient" only happened with single tests and was rare, suggesting such extremities are not a problem in practice using COTAN rules and given the sample sizes typically used. These results demonstrate that interpreting sample reliability values without taking CIs for population values into consideration may produce conclusions, which are too optimistic. We hope this study is a wake-up call for anyone involved in test construction and test assessment not to treat sample results as parameters, and to assess reliability using CIs allowing a statistically well-founded decision whether $\rho \geq c$ and a precise estimate of ρ. Power and precision may not be accomplished simultaneously (e.g., high power may go together with low precision reflected by wide CIs), but for several statistical procedures (not including reliability) Maxwell et al. (2008) discuss sample size planning aimed at obtaining both power and precision.

Should using hard category boundaries such as c_{IS} and c_{SG} be preferred to soft interval boundaries? Soft boundaries allow labeling .79, say, "rather good" if .80 being only .01 unit higher was labeled "good," whereas hard boundaries such as used by COTAN simply label .79 "insufficient" and .80 "good." We make two remarks. First, whatever categorization one uses, if the purpose is to classify tests, in the end one needs to make a decision based on numerical sample values for which we recommend using CIs that reflect ones uncertainty due to sample size. When samples are large, CIs are not important anymore and human judgment should be used to assess what is reasonable and may be inspired by considerations such as the uniqueness of the test and the sample available, hence the difficulty to replace either. For example, a braille intelligence test for blind people may be unique, hence impossible to replace, and even a small sample of blind people may be hard to obtain, so that one must use whatever data are available and accept results for use provided they are not disastrous. Second, to link reliability values more tightly to labels that most people agree about needs the introduction of external criteria with respect to test utility, for example, referring to numbers of false positives and false negatives. Relating test results to utility of outcomes is a complex topic that is both important and beyond the scope of this study.

Will categorization systems other than COTAN's produce different results? Probably, for example, if the systems' boundaries do not match the database's reliability values (boundaries are distant from where most reliability values are), or when a finer-grained system of boundaries is used so that intervals are narrower and more tests are downgraded more than one category. Related thereto, we also broke down results to test types and use scenarios corresponding to different boundary values, and found little differences for test types but found that as test use was more important (and boundaries higher), the percentage of reclassified tests decreased. For 95% CIs, we found for group-level test use 26.8% reclassification, for individual advice 20.5%, and for important individual decision 15.8%. A problem with these and similar breakdown results is that it is unknown whether trends like this would also be found with other databases and different categorization systems. We suspect these and similar results to be rather system-dependent and thus take such results not too literally.

Researchers and test constructors might consider using statistically more advanced methods for estimating CIs for coefficient alpha and the split-half method. Kelley and Cheng (2012) suggested a bootstrap method for obtaining CIs for methods other than split-half reliability and coefficient alpha. Cronbach (1951) intended alpha to replace the split-half method and since then many methods have been proposed that might be preferred to coefficient alpha because they are closer to true reliability ρ (Bentler & Woodward, 1980; Brennan, 2001; Guttman, 1945; Shavelson & Webb, 1991; Zinbarg et al., 2005). Hence, it is remarkable how persistent the use of coefficient alpha is, and even more the persistent albeit more modest use of the split-half method.

We recommend that test manuals and websites reporting test assessments standardize the information they provide comparable to a consumer's guide giving technical and user-relevant information for washing machines and dryers, cars, cell phones, and computers. However, the absence at the COTAN website of simple statistics like test length, sample size, and reliability estimates may often not be due not to lack of standardization of the website but perhaps more to the absence of this kind of information in test manuals. A hypothesis the authors discussed but were unable to check is that large testing agencies probably are better used to working according to protocol than researchers working in smaller companies, on their own or in small teams in hospitals, and also researchers working in a university environment where academic independence is highly valued, perhaps at the expense of standardization. It is extremely important that large testing agencies and assessment authorities such as COTAN emphasize that anyone constructing, publishing, and selling tests provides the relevant information about test quality.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143. doi:10.1007/s11336-008-9100-1

Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*, 249-267. doi:10.1007/BF02294079

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295-317. doi:10.1111/j.1745-3984.2001.tb01129

Cascio, W. F. (1991). *Applied psychology in personnel management*. Englewood Cliffs, NJ: Prentice-Hall.

Charter, R. A. (2000). Confidence interval formulas for split-half reliability coefficients. *Psychological Reports*, *86*, 1168-1170. doi:10.2466/PR0.86.3.1168-1170

Christensen, L. B. (1997). *Experimental methodology*. Boston, MA: Allyn & Bacon.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309-319. doi:10.1037/1040-3590.7.3.309

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.527.7772&rep=rep1&type=pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi:10.1007/BF02310555

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391-418. doi:10.1177/0013164404266386

DeVellis, R. F. (2003). *Scale development. Theory and applications*. Thousand Oaks, CA: Sage.

Egberink, I. J. L., Janssen, N. A. M., & Vermeulen, C. S. M. (2009-2016). *COTAN Documentatie* [COTAN documentation]. Amsterdam, Netherlands: Boom. Retrieved from https://www.cotandocumentatie.nl/

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests. Test review form and notes for reviewers. Version 4.2.6*. Retrieved from www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4

Evers, A. V. A. M., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2010). *COTAN beoordelingssysteem voor de kwaliteit van tests* [COTAN assessment system for quality of tests]. Retrieved from https://www.psynip.nl/wp-content/uploads/2016/07/COTAN-Beoordelingssysteem-2010.pdf

Evers, A. V. A. M., Sijtsma, K., Meijer, R. R., & Lucassen, W. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. *International Journal of Testing*, *10*, 295-317. doi:10.1080/15305058.2010.518325

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes, confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517-531. doi:10.1177/0013164401614001

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, *30*, 357-370. doi:10.1007/BF02289499

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93-103. doi:10.1177/014662168701100107

Gregory, R. J. (2007). *Psychological testing. History, principles and applications*. Boston, MA: Allyn & Bacon.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282. doi:10.1007/BF02288892

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika*, *41*, 219-231. doi:10.1007/BF02291840

Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Holt, Rinehart, & Winston.

Heiser, W., Hubert, L., Kiers, H., Köhn, H.-F., Lewis, C., Meulman, J., . . . Takane, Y. (2016). Commentaries on the ten most highly cited *Psychometrika* articles from 1936 to the present. *Psychometrika*, *81*, 1177-1211. doi:10.1007/s11336-016-9540-y

Kelley, K., & Cheng, Y. (2012). Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology*, *8*, 39-50. doi:10.1027/1614-2241/a000036

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, *28*, 221-238. doi:10.1007/BF02289571

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology*, *66*, 503-520. doi:10.1111/bmsp.12010

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Markon, K. E., & Chmielewski, M. (2013). The effect of response model misspecification and uncertainty on the psychometric properties of estimates. In R. E. Millsap, L. A. van der Ark, D. M. Bolt & C. M. Woods (Eds.), *New developments in quantitative psychology. Presentations from the 77th Annual Psychometric Society Meeting* (pp. 85-114). New York, NY: Springer. doi:10.1007/978-1-4614-9348-8_7

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537-563. doi:10.1146/annurev.psych.59.103006.093735

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157-176. doi:10.1037/1082-989X.12.2.157

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293-299. doi:10.1037/1082-989X.1.3.293

McIntire, S. A., & Miller, L. A. (2007), *Foundations of psychological testing. A practical approach*. Thousand Oaks, CA: Sage.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing. Principles and applications*. Upper Saddle River, NJ: Prentice-Hall.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13. doi:10.1007/BF02289400

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman's λ_4, λ_5, and λ_6 and the greatest lower bound. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas & W.-C. Wang (Eds.), *Quantitative psychology: The 81th Annual Meeting of the Psychometric Society 2016, Asheville NC* (pp. 159-172). New York, NY: Springer.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353. doi:10.1037/1040-3590.8.4.350

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory. A primer*. Thousand Oaks, CA: Sage.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120. doi:10.1177/0959354312454353

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, *77*, 4-20. doi:10.1007/s11336-011-9242-4

Sijtsma, K., & Van der Ark, L. A. (2015). Conceptions of reliability revisited and practical recommendations. *Nursing Research*, *64*, 128-136. doi:10.1097/NNR.0000000000000077

Smith, M. J., & Smith, P. (2005). *Testing people at work: Competencies in psychometric testing*. Malden, MA: Blackwell.

Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*, 575-579. doi:10.1007/BF02293815

Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271-280. doi:10.1007/BF02296146

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604. Retrieved from https://www.apa.org/science/leadership/bsa/statistical/tfsi-followup-report.pdf

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 122-133. doi:10.1007/s11336-003-0974-7